UDC 681.2:621.3.082.1 DOI:

Automated system for identification of data distribution laws by analysis of histogram proximity with sample reduction

O. Oliynyk [1], Yu. Taranenko [2]

[1] *College of Radio Electronics, Shmidta Str., 18, 49000, Dnipro, Ukraine*
oleinik_o@ukr.net
[2] *Private Enterprise "Likopak", Kachalova Str., 1, 49005, Dnipro, Ukraine*

Abstract

The error in the identification of the distribution law entails an incorrect assessment of other characteristics (standard deviation, kurtosis, antikurtosis, etc.). The article is devoted to the development of accessible and simple software products for solving problems of identifying distribution laws and determining the optimal size of a data sample.

The paper describes a modified method for identifying the law of data distribution by visual analysis of the proximity of histograms with a reduction in the sample size with software implementation. The method allows choosing the most probable distribution law from a wide base of the set. The essence of the method consists in calculating the entropy coefficient and absolute entropy error for the initial and half data sample, determining the optimal method for processing the histogram using visual analysis of the proximity of histograms, and identifying the data distribution law. The experimental data processing model makes it possible to take into account the statistical properties of real data and can be applied to various arrays, and allows to reduce the sample size required for analysis.

An automated system for identifying the laws of data distribution with a simple and intuitive interface has been developed. The results of the study on real data indicate an increase in the reliability of the identification of the data distribution law.

Keywords: reliability; distribution law; sample; entropy coefficient; histogram.

Received: 27.04.2021 Edited: 30.06.2021 Approved for publication: 06.07.2021

Introduction

The choice of the mathematical method of data processing, as well as the accuracy of obtaining numerical values by the metrological characteristics of measuring instruments to a significant extent depend on the correctness of establishing the correspondence of the probability distribution density of experimental data to one of the standard distribution laws [1]. The most common way to improve the reliability of the identification of a distribution is to improve the procedure for correlating the obtained data with one of the standard distributions.

Statistical processing of experimental data begins with calculating the center of distribution. Therefore, the error in the identification of the distribution law entails an incorrect assessment of the second characteristics (standard deviation, kurtosis, antikurtosis, etc.) [2]. Since obtaining a priori information about possible statistical characteristics of experimental data is very difficult (noisy data), increasing the reliability of identification of distribution laws is still an urgent metrological problem [1].

Analysis of recent research and publications

The diversity of forms of distribution of measurement errors is confirmed by both numerous publications [3–6] and the availability of standardized models of distribution laws [7]. The algorithms for processing experimental data are quite diverse [3–6]. The choice of the required algorithm for a specific measurement problem causes significant difficulties due to the difficulty of assessing the efficiency of experimental data processing.

In the theory of signal processing, a classical two-stage approach is used [1]. The essence of the algorithm comes down to constructing a histogram of the unknown distribution law of the error, determining the quantitative features of the measure of difference from the normal law (the coefficients of asymmetry and kurtosis were used in the work). Comparison of distribution laws: standard, experimental in the form of a histogram, polygon, approximating function, built on one graph, allows to establish the difference or similarity between them [1].

One of the most promising approaches to determining the distribution laws is the characterization of the shape of the distribution law using the antikurtosis ($\varkappa$) and the entropy coefficient (k) [6, 8–11]. According to information theory [8], for all possible existing distribution laws, the antikurtosis value lies in the range from 0 to 1, and k – from 0 to 2.076. In a number of works, the identification of distribution laws is considered in ($\varkappa$, k) – a plane in which each law is identified by some point [9–12]. This allows performing preliminary data processing with the definition of the distribution law for Kalman filtering [9], processing and decoding of cardiograms [10]. However, after the identification of the distribution law was performed, the studies did not assess the closeness of the histograms to the selected distribution; therefore, there is no data on the effectiveness of such techniques.

To implement the described algorithms, a large number of software products are used (for example, Mathematica, MATLAB, etc.). The widely used MathCAD [1, 12] allows to use the built-in ones for plotting histograms, polygons and approximating functions. The complexity of the selection of the required law based on the histogram, the time spent on enumerating all possible options force researchers to widely use computer binning. In general, binning is a preprocessing technique used to reduce the impact of minor observation errors. The original data values that fall within a given small interval, called a bin, are replaced by a value representing that interval, often a center value [13, 14]. When analyzing histograms, the essence of binning procedures is reduced to the selection of characteristics of the histogram cell [13, 14]. The use of software binning allows to get away from the problem of choosing a specific formula for the distribution law to the selection of the nearest identical data sample.

In publications [14, 15] binning models are presented and the main set of samples with characteristic features are considered; the identification problem is reduced to finding a close histogram with a known distribution law. The main disadvantage of this approach is the lack of existing sample models for assessing various samples of experimental data and the absence of a universal criterion for assessing the proximity of histograms negatively affects the identification error of the distribution law according to the statistical model.

However, the key moment that determines the efficiency of identifying the shape of distribution of experimental data is the sample size and the choice of the number of intervals for the distribution of data. In real conditions, ensuring the required sample size (more than 1000) is a difficult task associated, first of all, with the economic aspect of multiple measurements. In [12], it was proposed to use the entropy coefficient to choose the number of intervals for grouping data, the results were confirmed by the study of several distribution laws, which does not allow to speak about the universality of the proposed method.

Analysis of modern trends in data processing methods indicates the priority development of software methods for processing measurement information. The lack of accessible and simple software products for solving problems of identifying distribution laws and determining the optimal size of a data sample poses the task of researchers developing an automated system for identifying distribution laws taking into account the sample size.

Purpose and objectives of the study

The aim of the work is to develop a system for identifying the distribution laws of the sample data by visual analysis of the proximity of histograms to the real sample and reducing the sample size.

The optimal algorithm for constructing a histogram for a statistical model for determining the distribution law with the possibility of reducing the data sample

Studies of the proximity of histograms were carried out programmatically in Python using the numpy. histogram function, which calculates the boundaries of intervals for histogram cells and has a fairly wide range of methods for calculating the optimal width of histogram cells [15, 16].

To study the proximity of histograms, we used real arrays of experimental data [17]. The algorithm for estimating the proximity of histograms begins with calculating the absolute entropy error [8] and the entropy coefficient.

Absolute entropy error [8]:

$$h = 0.5 * e^{H(x)} = 0.5 \cdot d \cdot n \cdot 10^{-\frac{1}{n}\sum_{1}^{m} n_i \cdot \log(n_1)}, \qquad (1)$$

where $H(x)$ – the entropy of the distribution; n is the sample size, d – the width of the histogram column; m is the number of histogram columns, n_i – the number of observations in the m column.

The entropy coefficient is defined [6, 8, 10] as:

$$k_e = \frac{h}{\sigma}, \qquad (2)$$

where σ – the standard deviation; h – the entropy value of the error.

To analyze histograms, one of the methods for calculating the optimal width of histogram cells ("auto", "fd", "doane", "scott", "stone", "rice", "sturges", "sqrt") is used [15, 16]. For all these methods for calculating the optimal width of the histogram cells, we determine the minimum difference between the entropy coefficients for the original sample (k_e) and for the sample obtained from the original by excluding odd terms ($k_{0.5e}$), according to the ratio:

$$z = \min(k_e - k_{0.5e}), \qquad (3)$$

where z is the decrease in entropy for a sample that is two times smaller.

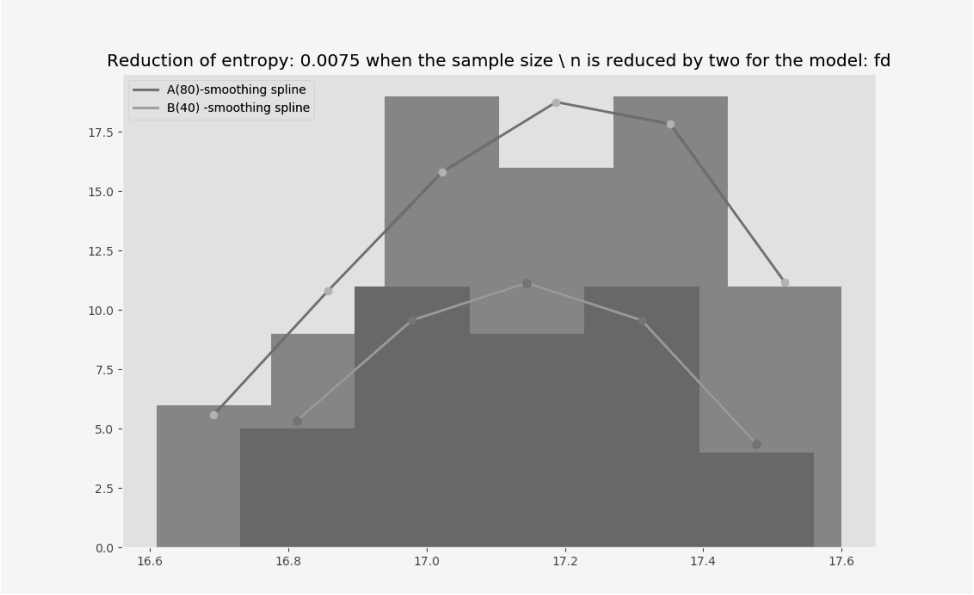

Fig. 1. The result of determining the proximity of histograms using the minimum difference z of the entropy coefficients for the original sample (k_e) and half sample ($k_{0.5e}$) for the method for calculating the optimal width of histogram cells (model "fd" – cubic smoothing spline)

An example of the results of determining the proximity of histograms using a cubic smoothing spline is shown in Fig. 1. In accordance with the described algorithm, a histogram was obtained that optimally displays the original data sample when it is reduced by 2 times.

This approach makes it possible to rationally reduce the amount of required data and the number of multiple measurements while maintaining the data distribution law, which reduces the cost of metrological certification of measuring instruments. To confirm the effectiveness of the developed method and compare, we present the results of the algorithm when calculating the maximum difference in entropy coefficients by the expression:

$$z = \max(k_e - k_{0.5e}). \quad (4)$$

In Fig. 2, it is noticeable that with a similar reduction in the sample, the difference in entropy coefficients is significantly higher than when using (3), and, accordingly, the proximity of the histograms is less.

The list of distribution laws available for analysis is wide and includes about 80 distributions and criteria for comparison (the most common law, distribution density range, desired identification error).

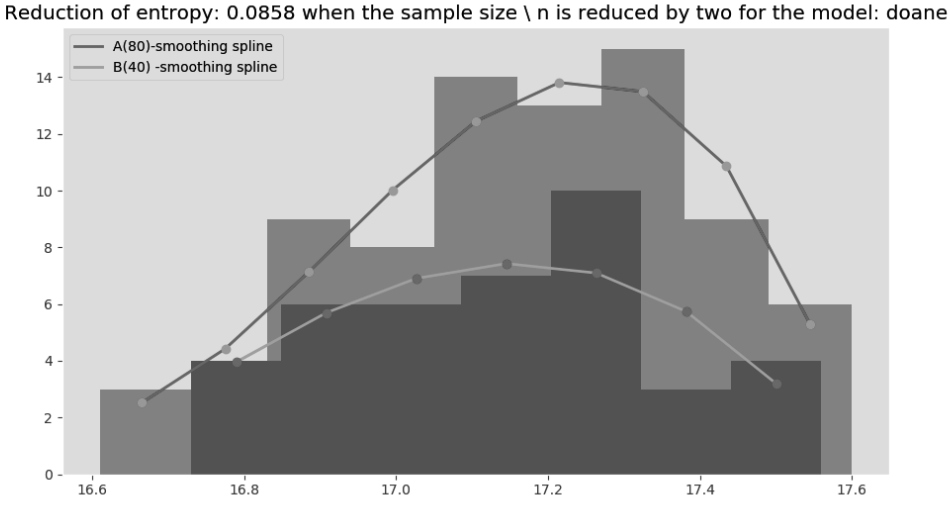

Fig. 2. The result of determining the proximity of histograms using the maximum difference z of the entropy coefficients for the original sample (k_e) and a half sample ($k_{0.5e}$) for the method for calculating the optimal width of histogram cells ("doane" model – for data with an abnormal distribution of values)

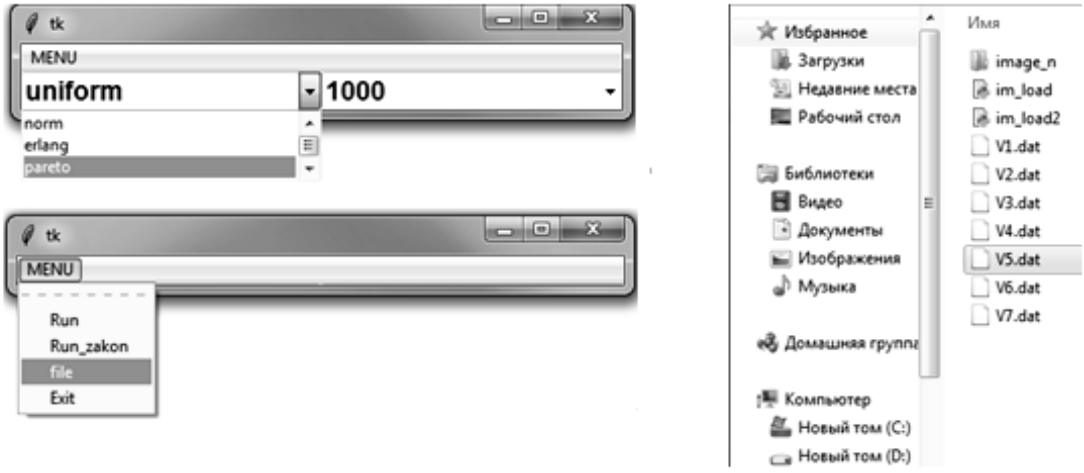

Fig. 3. Interface of the automated system for identifying the laws of data distribution

The coincidence of the histograms indicates the maximum closeness to this distribution law.

Fig. 3 shows the interface of the developed automated system for identifying the laws of data distribution.

Work in the program begins with the choice of loading a file with data sampling (Fig. 3) and using the command "Menu", "file". Next, the system visualizes the histograms of the analyzed sample, comparing the histograms for full and half samples of real data, according to the described algorithm using various models of methods for calculating the optimal width of histogram cells (Fig. 4).

The last step is to identify the distribution law for sampling real measurement data (Fig. 5).

The identification error of the distribution law for the selected sample $n = 1000$ does not exceed 0.45%. According to [18], a convincing advantage for use is the accuracy of the method for identifying the distribution law above 3%.

The developed automated system for data processing, designed to identify the laws of data distribution by visual analysis of the proximity of histograms, can be used to process measurement information during metrological certification of measuring instruments. The program has a flexible model for processing experimental data and takes into account the statistical properties of real data. This allows the automated system to be used to solve a wide range of measurement tasks.

Conclusions

1. A modified method for identifying the law of data distribution by visual analysis of the proximity of histograms is proposed. The method allows choosing the most probable distribution law from a wide base of the set. The essence of the method consists in calculating the entropy coefficient and absolute entropy error for the initial and half sample of data, determining the optimal method for processing the

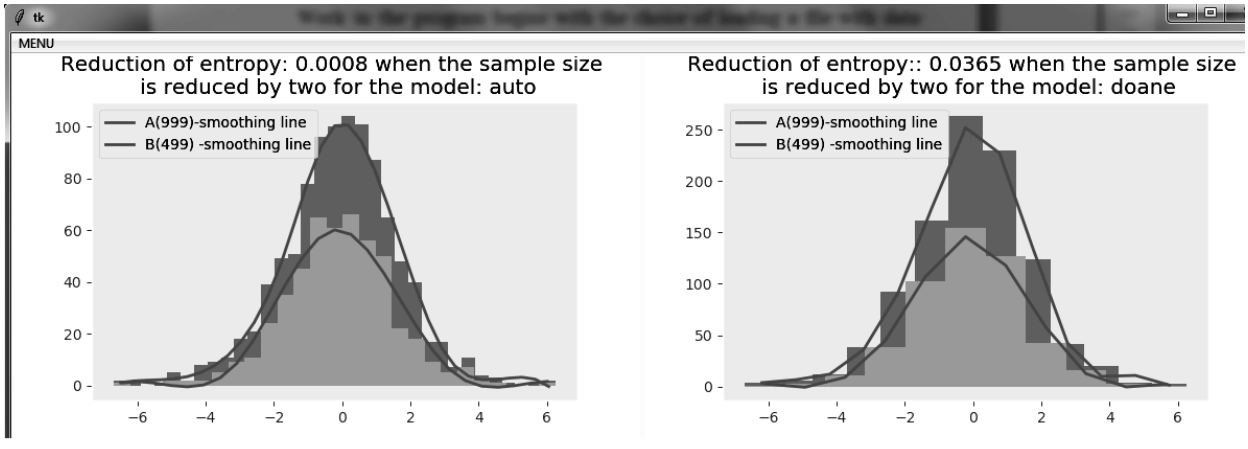

Fig. 4. Determination of the method for calculating the optimal width of the histogram cells ("auto", "fd", "doane", "scott", "stone", "rice", "sturges", "sqrt") while minimizing the entropy coefficient

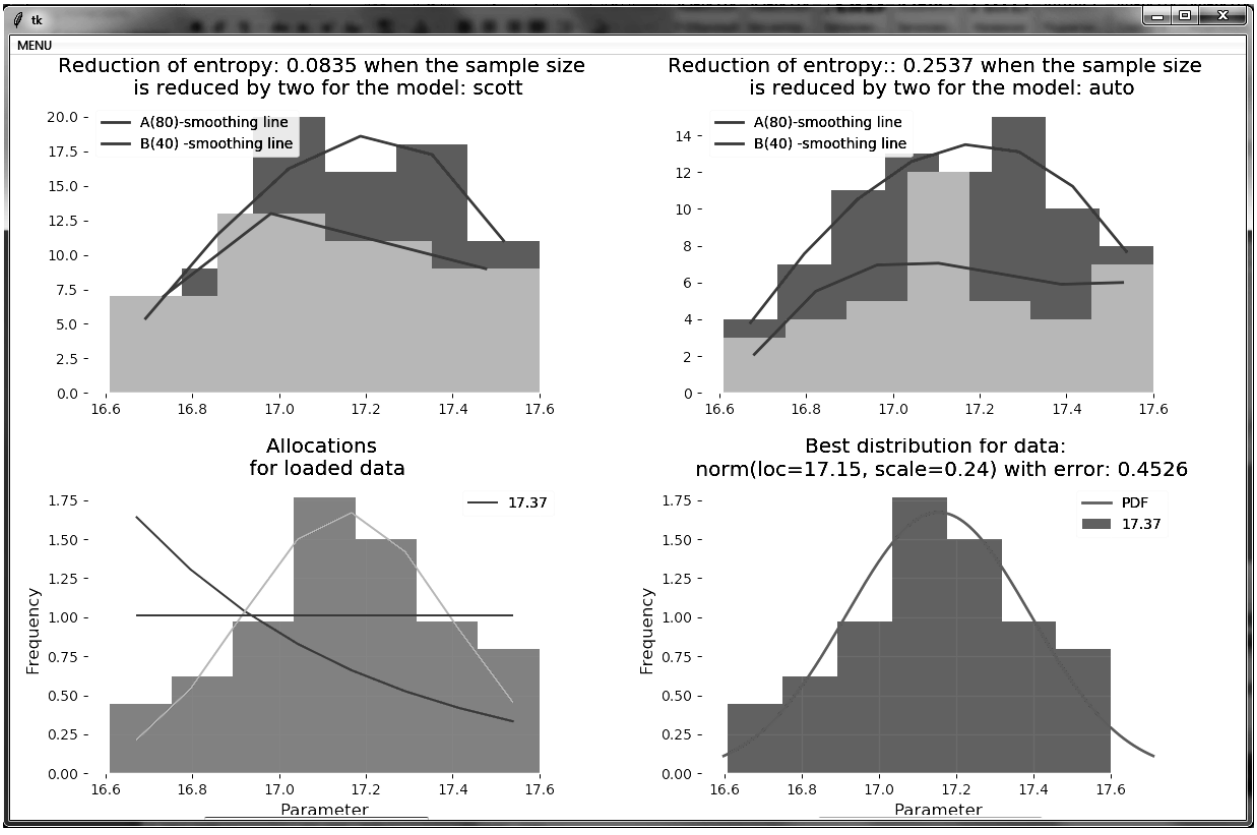

Fig. 5. The results of the automated system for sampling n = 1000 from the uniform distribution

histogram using visual analysis of the proximity of histograms, and identifying the law of data distribution.

2. The advantages of the proposed method for identifying the distribution law are simplicity and flexibility of the experimental data processing model, which makes it possible to take into account the statistical properties of real data and can be applied to various data sets. It allows to reduce the sample of data required for analysis.

3. On the basis of the described method, an automated system for identifying the laws of data distribution has been developed. The study carried out on experimental data indicates an increase in the reliability of identification of the data distribution law.

Автоматизована система ідентифікації законів розподілу даних аналізом близькості гістограм зі скороченням вибірки

О.Ю. Олійник[1], Ю.К. Тараненко[2]

[1] *Коледж радіоелектроніки, вул. Шмідта, 18, 49000, Дніпро, Україна*
oleinik_o@ukr.net
[2] *ПП "Лікопак", вул. Качалова, 1, 49005, Дніпро, Україна*

Анотація

Похибка визначення закону розподілу тягне за собою невірну оцінку інших характеристик (стандартне відхилення, ексцес, контрексцес і т.д.). Точність отримання числових значень метрологічних характеристик засобів вимірювальної техніки значною мірою залежить від правильності встановлення відповідності щільності розподілу ймовірностей експериментальних даних одному зі стандартних законів розподілу. Однак після виконаної ідентифікації закону розподілу в дослідженнях не оцінювалася близькість гістограм заданого розподілу, тому і даних про ефективність таких методик немає. Статтю присвячено розробці доступних і простих програмних продуктів для розв'язання задач виявлення законів розподілу і визначення оптимального розміру вибірки даних. Розроблено модифікований метод визначення закону розподілу даних шляхом візуального аналізу близькості гістограм зі зменшенням розміру вибірки при програмній реалізації. Метод дозволяє вибрати найбільш імовірний закон розподілу

з широкої бази набору. Суть методу полягає в обчисленні ентропії коефіцієнта і абсолютної ентропійної помилки для вихідної та половинної вибірки даних, визначенні оптимального методу обробки гістограми за допомогою візуального аналізу близькості гістограм і виявленні закону розподілу даних. Модель обробки експериментальних даних дозволяє враховувати статистичні властивості реальних даних і може застосовуватися до різних масивів, а також дозволяє зменшити розмір вибірки, необхідної для аналізу. Розроблено автоматизовану систему визначення законів розподілу даних із простим і зрозумілим інтерфейсом. Результати дослідження на реальних даних свідчать про підвищення достовірності ідентифікації закону розподілу даних. Похибка ідентифікації закону розподілу з використанням розробленого методу для вибірки n = 1000 не перевищує 0,45% у порівнянні з точністю 3% для відомих методів.

Ключові слова: достовірність; закон розподілу; вибірка; ентропійний коефіцієнт; гістограма.

Автоматизированная система идентификации законов распределения данных анализом близости гистограмм с сокращением выборки

О.Ю. Олейник[1], Ю.К. Тараненко[2]

[1] *Колледж радиоэлектроники, ул. Шмидта, 18, 49000, Днепр, Украина*
oleinik_o@ukr.net
[2] *ЧП "Ликопак", ул. Качалова, 1, 49005, Днепр, Украина*

Аннотация

Погрешность определения закона распределения влечет за собой неверную оценку других метрологических характеристик. Статья посвящена разработке программных продуктов для решения задач выявления законов распределения и определения оптимального размера выборки данных.

Авторами описан модифицированный метод определения закона распределения данных путем визуального анализа близости гистограмм с уменьшением размера выборки при программной реализации. Суть метода заключается в вычислении энтропийного коэффициента и абсолютной энтропийной ошибки для исходной и половинной выборки данных, определении оптимального метода обработки гистограммы с помощью визуального анализа близости гистограмм и выявлении закона распределения данных. Модель обработки экспериментальных данных позволяет учитывать статистические свойства реальных результатов и может применяться к различным массивам. Разработана автоматизированная система определения законов распределения данных с простым и понятным интерфейсом. Результаты исследования на реальных данных свидетельствуют о повышении достоверности идентификации закона распределения данных.

Ключевые слова: достоверность; закон распределения; выборка; энтропийный коэффициент; гистограмма.

References

1. Zyuzko A.K., Burichenko M.Yu., Petrova Yu.V., Nimych V.V. Algorithm of treatment of data given during metrology attestation of facilities. *Electronics and Control Systems*, 2009, no. 1, pp. 5–10 (in Russian).
2. Shchepetov A.G., Pidkovich A.A., Popova Ya.D., Shimereva L.V. O vybore metoda obrabotki eksperimentalnykh dannykh pri opredelenii graduirovochnoy staticheskoy kharakteristiki pribora [On the choice of the experimental data processing method when determining the calibration static characteristics of the device]. *Instruments*, 2020, no. 9, pp. 14–22 (in Russian).
3. Bityukov S.I., Maksimushkina A.V., Smirnova V.V. Comparison of histograms in physical research. *Izvestiya vuzov. Yadernaya Energetika*, 2016, no. 1, pp. 81–90 (in Russian).
4. Hou J., Ou B., Tian H., Qin Z. Reversible data hiding based on multiple histograms modification and deep neural networks. *Signal Processing: Image Communication*, 2021, vol. 92, pp. 116–118.
5. Liang Y., Meng Z., Chen Y., Zhang Y., Wang M., Zhou X. A Data Fusion Orientation Algorithm Based on the Weighted Histogram Statistics for Vector Hydrophone Vertical Array. *Sensors*, 2020, vol. 20, no. 19, p. 5619. doi: https://doi.org/10.3390/s20195619

6. Artyushenko V.M., Volovach V.I. Identifikatsiya parametrov raspredeleniya additivnykh i multiplikativnykh negaussovskikh pomekh [Identification of distribution parameters of additive and multiplicative non-Gaussian noise]. *Avtometriya*, 2017, vol. 53, no. 3, pp. 36–43 (in Russian). doi: 10.15372/AUT20170305
7. MI 1317-2004. GSE. Results and characteristics of measurement errors. Forms of presentation. Methods of use when testing product samples and controlling their parameters. Moscow, 2004 (in Russian).
8. Novitsky P.V., Zograf I.A. Otsenka pogreshnostey rezultatov izmereniy [Estimation of errors of measurement results]. Leningrad, Energoatomizdat Publ., 1991 (in Russian).
9. Oliynyk O., Taranenko Y., Losikhin D., Shvachka A. Examining the Kalman filter in the field of noise and interference with the Non-Gaussian distribution. *Eastern-European Journal of Enterprise Technologies*, 2018, vol. 4, no. 4(94), pp. 36–42. doi:10.15587/1729-4061.2018.140649
10. Bodin O.N., Ivanchukov A.G., Polosin V.G., Rahmatullov F.K. Entropiyno-parametricheskaya obrabotka elektrokardiosignala [Entropy-parametric processing of electrocardiosignal]. *Fundamental research*, 2015, no. 3, pp. 23–27 (in Russian).
11. Tynynyka A.N. Primeneniye entropiynogo koeffitsiyenta dlya optimizatsii chisla intervalov pri intervalnykh otsenkakh [Application of the entropy coefficient for optimization of the number of intervals in interval estimates]. *Tekhnologiya i Konstruirovanie v Elektronnoi Apparature*, 2017, no. 3, pp. 49–54 (in Russian). doi: 10.15222/TKEA2017.3.49
12. Fedorov M.V. Metod identifikatsii form raspredeleniy malykh vyborok [Method of identification of forms of distributions of small samples]. *Rossijskij himicheskij zhurnal*, 2002, no. 3, pp. 9–11 (in Russian).
13. Python histogram. Python rendering matplotlib13. Histogram (histogram) detailed explanation. Available at: https.blog.csdn.net/weixin_39520979/article/details/111293856 (accessed 12.04.2021)
14. Sulewski P. Equal-bin-width histogram versus equal-bin-count histogram. *Journal of Applied Statistics*, 2020, pp. 1–20. doi: https://doi.org/10.1080/02664763.2020.1784853
15. Reducing the sample size of experimental data without losing information. Available at: https://habr.com/ru/post/445464/ (accessed 05.25.2019).
16. Numpy.histogram_bin_edges. Available at: https://numpy.org/doc/stable/reference/generated/numpy.histogram_bin_edges.html (accessed 05.25.2019).
17. Kalmykov V.V., Antonyuk F.I., Zenkin N.V., Malyshev E.N. Organizatsiya statisticheskogo priyemochnogo kontrolya kachestva produktsii po kolichestvennomu priznaku [Organization a statistical acceptance inspection quality products at the quantitative trait]. *Modern Problems of Science and Education*, 2014, no. 6, p. 162 (in Russian).
18. Tyrsin A.N. Metod podbora nailuchshego zakona raspredeleniya nepreryvnoy sluchaynoy velichiny na osnove obratnogo otobrazheniya [The method of selecting the best distribution law for continuous random variables on the basis of inverse mapping]. *Vestn. Yuzhno-Ural. Gos. Un-ta. Ser. Matem. Mekh. Fiz.*, 2017, vol. 9, no. 1, pp. 31–38 (in Russian).